Extreme change-point detection

Kevin Bleakley[1,2,3]

[1]LMO, Orsay
[2]Inria
[3]CNRS

Abstract

We examine rules for predicting whether a point in $\mathbb{R}$ generated from a 50–50 mixture of two different probability distributions came from one distribution or the other, given limited (or no) information on the two distributions, and—as clues—one point generated randomly from each of the two distributions. We prove that nearest-neighbor prediction does better than chance when we know the two distributions are Gaussian densities without knowing their parameter values. We conjecture that this result holds for general probability distributions and—furthermore—that the nearest-neighbor rule is optimal in this setting, i.e., no other rule can do better than it if we do not know the distributions or do not know their parameters, or both.

I. Introduction

This work originated in trying understand—in the most simple setting possible—what detecting one change-point in a time series truly means. Suppose we know that there is one (and exactly one) change-point in a real-valued time series of length n, located between the kth and $k+1$th data points; i.e., the first k points are randomly drawn from some probability distribution f_X, and the last $n-k$ points from another, different, probability distribution f_Z. The minimal interesting setting is when $n = 3$. In it, we know that the first point X is generated from f_X, the third point Z from f_Z, and it remains to try to work out which of the two distributions the middle point Y came from (which is equivalent to predicting the change-point location).

If we are told that the middle point is more likely to have been drawn from one of the two distributions, we can already obtain a decision rule that is better than chance: *Always predict the more likely distribution, no matter the values of the three points*. However, with real-world data, we are unlikely to have the slightest clue about which distribution the middle point comes from. In this case, it makes sense that in the absence of prior knowledge, we treat the middle point as if it were generated from a 50–50 mixture distribution of the two distributions: $Y \sim \frac{1}{2}f_X + \frac{1}{2}f_Z$.

Note that in this setting, if we had full knowledge of the distributions f_X and f_Z, knowing the values of the random draws $X = x$ and $Z = z$ would not in fact provide us with additional information as to whether $Y = y$ was generated from f_X or f_Z; indeed, in this case we already have all the information we need to calculate the (optimal) Bayes classifier, (given $Y = y$), that is, the classifier that minimizes the probability of incorrect prediction of the distribution y came from. Again however, in the real world with real data, we are quite unlikely to know the distributions f_X and f_Z a priori. That said, in what follows, we only deal with the $n = 3$ case, and these results do not generalize to $n > 3$ or $\mathbb{R}^d$ for $d > 1$. Nevertheless, the results are interesting in their own right, given that they seem to be highlighting curious connections between probability distributions and distances between points from these distributions.

We are interested in two distinct issues: (1) proving that there exists some decision rule that is better than a coin flip in settings where we have little or no information on the two probability distributions, and (2) proving that this decision rule is optimal, i.e., that no other prediction rule has a lower probability of incorrect prediction. In the following, we have partial answers to (1), but (2) remains completely unanswered. It may seem intuitive to some that a nearest-neighbor rule cannot be beaten in $\mathbb{R}$, but it is another thing to prove it.

II. A general conjecture

Let us set the scene with a general conjecture.

Conjecture II.1 *Suppose that x and z are drawn from arbitrary unknown probability distributions f_X and f_Z which are different in the sense that*

$$\int_{\mathbb{R}} |f_X(w) - f_Z(w)| dw > 0.$$

(In some other sense could be possible too.) Suppose also that y is drawn from the mixture model $Y \sim \frac{1}{2}f_X + \frac{1}{2}f_Z$. Let the decision rule for deciding whether y was drawn from f_X or f_Y be the nearest neighbor rule, i.e., predict that y is drawn from the distribution f_X if $|x - y| < |z - y|$, and vice versa. Then (i) this rule is correct more than half the time, i.e., better than a coin flip, and (2) this rule minimizes the classification error, i.e., no other rule does better.

Remark II.1 *This conjecture, if true, means that knowing only that f_X and f_Z are different (in the above sense), we can correctly predict the distribution y came from more than half the time.*

This conjecture turns out to be non-trivial, even if we suppose that the two unknown distributions are also (unknown) probability densities. Hence, in order to take tentative steps forward, we will initially relax the problem to settings where we at least know that the two distributions are probability densities, even if we don't know their precise forms (parameters, etc.).

III. The Gaussian setting with equal variance

Suppose that we are lucky and know that f_X and f_Z are Gaussian distributions with the same variance, but have no information on the value of this variance, nor on that of the two means, except what little can be gleaned from the three numbers x, y, and z. Formally, suppose that $X \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Z \sim \mathcal{N}(\mu_Z, \sigma^2)$. In this Gaussian setting, to simplify notation and proofs, we shall write ϕ_{m,σ^2} to mean the Gaussian probability distribution function (pdf) with mean m and variance σ^2, and Φ_{m,σ^2} the corresponding Gaussian cumulative distribution function (cdf).

Given x, y, and z drawn as before, we can ask questions like, "Is there a decision rule for predicting ϕ_{μ_X,σ^2} or ϕ_{μ_Z,σ^2} for y that is better than flipping a coin?", "Is there an optimal decision rule given what we know and don't know about the distributions and given x, y, and z?", and, "How close can we get to the optimal decision rule that would be known if we knew the true values of μ_X, μ_Z, and σ^2?"

I. The Bayes classifier in this setting

If we do know μ_X, μ_Z, and σ^2, the Bayes classifier C^{Bayes}, i.e., the rule that minimizes the probability of error when predicting ϕ_{μ_X,σ^2} vs ϕ_{μ_Z,σ^2}, is easy to calculate:

$$C^{Bayes}(y) = \begin{cases} \phi_{\mu_X,\sigma^2} & \text{if } \phi_{\mu_X,\sigma^2}(y) > \phi_{\mu_Z,\sigma^2}(y) \\ \phi_{\mu_Z,\sigma^2} & \text{if } \phi_{\mu_X,\sigma^2}(y) < \phi_{\mu_Z,\sigma^2}(y), \end{cases} \tag{1}$$

i.e., predict the distribution of y as that which has the highest density at y.

Remark III.1 *Note that when both Gaussian distributions have the same variance, as is the case here, the Bayes classifier is equivalent to the rule:*

$$C^{\mu}(y) := \begin{cases} \phi_{\mu_X,\sigma^2} & \text{if } |y-\mu_X| < |y-\mu_Z| \\ \phi_{\mu_Z,\sigma^2} & \text{if } |y-\mu_X| > |y-\mu_Z|, \end{cases} \tag{2}$$

where $|\cdot|$ is the absolute value. i.e., predict the distribution whose mean is closest to y. Note that though this is not exactly a nearest-neighbor rule, since μ_X and μ_Y are not data points, it is not too far off one.

II. Better than a coin flip

It turns out that knowing only that we have two Gaussian distributions with the same variance, and given x, y, and z drawn as before, there exist rules that are better than a coin flip if the (unknown) means μ_X and μ_Z are different. (If the two means are the same, then the two Gaussian distributions are identical and a coin flip is indeed the optimal rule.)

Theorem III.1 *Suppose that $X \sim \phi_{\mu_X,\sigma^2}$, $Z \sim \phi_{\mu_X+\epsilon,\sigma^2}$, and that Y is a 50–50 mixture of the two, where $\mu_X \in \mathbb{R}$, $\epsilon \neq 0$, and σ^2, are all unknown. Suppose we have x, z, and y generated respectively from these three distributions. Then the decision rule,*

$$C^{dist}(y) := \begin{cases} \phi_{\mu_X,\sigma^2} & \text{if } |x-y| < |z-y| \\ \phi_{\mu_X+\epsilon,\sigma^2} & \text{if } |x-y| > |z-y|, \end{cases} \tag{3}$$

has a probability of being correct greater than $1/2$.

Remark III.2 *We see that it is as if we are roughly estimating μ_X by x and $\mu_X + \epsilon$ by z and then using these point estimates of the means in the Bayes classifier.*

Proof. This decision rule will be correct when either y is closest to x and was drawn from f_X, or closest to z and was drawn from f_Z. Thus we want to prove that:

$$\frac{1}{2}\mathbb{P}(|X-Y| < |Z-Y| \mid Y \sim \phi_{\mu_X,\sigma^2}) + \frac{1}{2}\mathbb{P}(|X-Y| > |Z-Y| \mid Y \sim \phi_{\mu_X+\epsilon,\sigma^2}) > \frac{1}{2}. \tag{4}$$

Since the variance of ϕ_{μ_X,σ^2} and $\phi_{\mu_X+\epsilon,\sigma^2}$ is the same here, by symmetry it suffices to prove that

$$P^* := \mathbb{P}(|X-Y| < |Z-Y| \mid Y \sim \phi_{\mu_X,\sigma^2}) > 1/2.$$

Since the value of μ_X has no influence on the following calculations, we set $\mu_X = 0$ to simplify notation. Let us define the following function of x and z:

$$f(x,z) = \mathbb{P}(|x-Y| < |z-Y| \,|Y \sim \phi_{0,\sigma^2}). \tag{5}$$

The values of the random variable Y for which $|x-Y| < |z-Y|$ are those below $(x+z)/2$ if $x < z$, or those above $(x+z)/2$ if $x > z$. The measure of the set for which $|x-Y| < |z-Y|$ is therefore:

$$f(x,z) = \Phi_{0,\sigma^2}\left(\frac{x+z}{2}\right)\mathbb{1}_{\{x<z\}} + (1-\Phi_{0,\sigma^2})\left(\frac{x+z}{2}\right)\mathbb{1}_{\{x>z\}}. \tag{6}$$

Integrating $f(x,z)$ over x and z will give us the probability we are looking for:

$$P^* := P_1^* + P_2^* = \int_{x=-\infty}^{\infty}\int_{z=x}^{\infty}\Phi_{0,\sigma^2}\left(\frac{x+z}{2}\right)\phi_{\epsilon,\sigma^2}(z)\phi_{0,\sigma^2}(x)dzdx + \int_{x=-\infty}^{\infty}\int_{z=-\infty}^{x}\left(1-\Phi_{0,\sigma^2}\left(\frac{x+z}{2}\right)\right)\phi_{\epsilon,\sigma^2}(z)\phi_{0,\sigma^2}(x)dzdx.$$

Remarking that

$$\Phi_{0,\sigma^2}\left(\frac{x+z}{2}\right) = \Phi_{0,\sigma^2}\left(\frac{(x-\delta)+(z+\delta)}{2}\right)$$

for all $\delta \in \mathbb{R}$, the following change of variable greatly simplifies the integration:

- For P_1^* let $r = (x+z)/2$ and $\alpha = (z-x)/2$, i.e., $x = r-\alpha$ and $z = r+\alpha$.
- For P_2^* let $r = (x+z)/2$ and $\alpha = (x-z)/2$ i.e., $x = r+\alpha$ and $z = r-\alpha$.

In both cases, the absolute value of the Jacobian is 2. Let us concentrate on calculating the first double integral P_1^* above; the second—P_2^*—involves almost identical calculations. We have that:

$$P_1^* = 2\int_{r=-\infty}^{\infty}\int_{\alpha=0}^{\infty}\Phi_{0,\sigma^2}(r)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(r+\alpha-\epsilon)^2}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(r-\alpha)^2}d\alpha dr. \tag{7}$$

We consider the two Gaussian pdfs as if they were functions of α and use the fact that the product of two Gaussian pdfs is proportional to another Gaussian pdf to rewrite the double integral as:

$$P_1^* = \int_{r=-\infty}^{\infty}\Phi_{0,\sigma^2}(r)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma'^2}(r-\frac{\epsilon}{2})^2}\left(\int_{\alpha=0}^{\infty}\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{1}{2\sigma'^2}(\alpha-\frac{\epsilon}{2})^2}d\alpha\right)dr,$$

where $\sigma'^2 = \sigma^2/2$. The integral in α is nothing other than $1-\Phi_{\epsilon/2,\sigma'^2}(0)$, or equivalently, $\Phi_{-\epsilon/2,\sigma^2/2}(0)$ (which is independent of r). As for the integral in r, recalling the well-known result:

$$\int_{u=-\infty}^{\infty}\Phi_{0,1}\left(\frac{u-c}{\tau_1}\right)\phi_{0,1}\left(\frac{u-b}{\tau_2}\right)du = \tau_2\,\Phi_{0,1}\left(\frac{b-c}{\sqrt{\tau_1^2+\tau_2^2}}\right),$$

it simplifies to $\Phi_{0,1}\left(\frac{\epsilon}{\sqrt{6}\,\sigma}\right)$, and thus:

$$P_1^* = \Phi_{-\frac{\epsilon}{2},\frac{\sigma^2}{2}}(0)\Phi_{0,1}\left(\frac{\epsilon}{\sqrt{6}\,\sigma}\right).$$

Essentially identical calculations show that:

$$P_2^* = \Phi_{\frac{\epsilon}{2},\frac{\sigma^2}{2}}(0)\left(1-\Phi_{0,1}\left(\frac{\epsilon}{\sqrt{6}\,\sigma}\right)\right).$$

To begin to conclude, note that $\Phi_{-\epsilon/2,\sigma^2/2}(0) = \Phi_{0,1}(\epsilon/\sqrt{2}\,\sigma)$ and $\Phi_{\epsilon/2,\sigma^2/2}(0) = 1 - \Phi_{0,1}(\epsilon/\sqrt{2}\,\sigma)$, so that:

$$P_1^* + P_2^* = \Phi_{0,1}\left(\frac{\epsilon}{\sqrt{2}\,\sigma}\right)\Phi_{0,1}\left(\frac{\epsilon}{\sqrt{6}\,\sigma}\right) + \left(1 - \Phi_{0,1}\left(\frac{\epsilon}{\sqrt{2}\,\sigma}\right)\right)\left(1 - \Phi_{0,1}\left(\frac{\epsilon}{\sqrt{6}\,\sigma}\right)\right). \quad (8)$$

If $\epsilon > 0$, the domain of $\phi_{0,1}$ can be cut up into three pieces: $]-\infty, \epsilon/\sqrt{6}\,\sigma]$, $]\epsilon/\sqrt{6}\,\sigma, \epsilon/\sqrt{2}\,\sigma]$, and $]\epsilon/\sqrt{2}\,\sigma, \infty[$ with respective areas under $\phi_{0,1}$ of j, k, and ℓ. Thus $(j + k + \ell) = 1$. Hence, $P_1^* + P_2^* = (j + k)j + \ell(k + \ell) = j + \ell - 2j\ell$, using the fact that $(j + k + l)^2 = 1$ and expanding and rearranging. Then $P_1^* + P_2^* > 1/2$ is equivalent to $j > 1/2$ since $\epsilon > 0$ and $\ell < 1/2$. But $j > 1/2$ also since $\epsilon > 0$. A similar argument works when $\epsilon < 0$. □

Remark III.3 *It is easy to see that that as $\epsilon \to 0$, the two Gaussians become increasingly indistinguishable and this rule tends to a probability of being correct of $1/2$, from above. When $\epsilon \to \infty$, both this rule and the Bayes classifier tend to probability of being correct of 1. One could ask whether we could at least get an approximate idea of how well we might expect to do (between $1/2$ and 1) by roughly estimating ϵ by $z - x$, but we see in Eq. 8 that the value of $P_1^* + P_2^*$ obtained still depends very much on the (unknown) σ.*

Remark III.4 *One cannot help suspecting that Eq. 8 is related to products of areas under Gaussian densities connected to where they (or suitable normalized versions of them) cross over each other. For instance, the densities ϕ_{0,σ^2} and ϕ_{ϵ,σ^2} cross once at $x = \epsilon/2$ and their cdfs at this point are respectively equal to $\Phi_{0,1}(\epsilon/2\sigma)$ and $1 - \Phi_{0,1}(\epsilon/2\sigma)$.*

Given that we have only three data points, we may ask how this nearest neighbor classifier compares to other strategies such as maximum likelihood, CUSUM, and so on. We have the following corollary.

Corollary III.1 *Under the same conditions as Theorem III.1, decision rule in Eq. 3 corresponds to the same solution inferred from the maximum likelihood estimator, the CUSUM method, and linear and Gaussian kernels when running kernel change-point detection [1]).*

Proof. Without loss of generality, suppose that $x < z$ and $y < (x + z)/2$. If you run the calculations, the posterior of the maximum likelihood estimator (requiring the EM algorithm here) essentially says that it is more likely that y came from the same distribution as x than from the distribution of z, i.e., it picks the closest point to y as its classification rule. As for the CUSUM method (as described in [4]), some tedious algebra shows that the maximum absolute value of the CUSUM criterion for the two possible change-point locations is again equivalent to the rule predicting that y comes from the same distribution as the point (x or z) closest to it. As for kernel change-point detection, the linear kernel corresponds exactly to performing least-squares minimisation of a signal with piecewise constant mean and exactly one change-point, with the same variance in both constant segments. A few lines of algebra confirm that the change-point location that minimizes the sum of squared errors is again equivalent to predicting that y comes from the same distribution as the point (x or z) closest to it. Basic calculations show that the same is also true with the Gaussian kernel. □

Remark III.5 *Though tempted to conjecture that the same is true in general for other kernels, one quickly finds a counter-example: For the 1-d polynomial kernel of degree 2: $k_2(u, v) = (uv)^2$, if $x = 1$, $y = 2$, and $z = 2.9$, and thus y is closer to z than to x, the minimum of the kernel change-point criterion occurs when grouping x with y, i.e., when putting a change-point between y and z. This means that in some sense, in the eyes of the quadratic polynomial kernel, points further apart can be "more similar" to each other than points closer together.*

This brings us to the heart of our other question: Can we do better than the decision rule in Theorem III.1, given our hypotheses? Or is it optimal, and why? If we can do demonstrably better, is this new rule optimal? And if we cannot provide a better rule, how can we prove that the "nearest neighbor" rule is the best we can do, since we know it is not equal to the Bayes classifier? Is this finally a question of the geometry of 1-d space? We leave this as a conjecture.

Conjecture III.1 *Under the hypotheses of Theorem III.1, the optimal rule for deciding whether y came from the same distribution as x or z is the nearest neighbor rule defined in Theorem III.1.*

IV. The Gaussian setting with different variances

A natural question to ask after the previous section is whether this nearest neighbor decision rule is still valid if we know that the variances are—or could be—different for the two Gaussian distributions. Though intuition suggests that it is still valid, it turns out that certain steps in the proof of Theorem III.1 no longer work once the two variances are different. For instance, the symmetry argument whereby

$$P^* := \mathbb{P}(|X - Y| < |Z - Y| \mid Y \sim \phi_{\mu_X, \sigma^2}) > 1/2$$

implies

$$P^* := \mathbb{P}(|X - Y| > |Z - Y| \mid Y \sim \phi_{\mu_{X+\epsilon}, \sigma^2}) > 1/2$$

no longer holds in general; indeed, it turns out—surprisingly—to be possible that one of these two terms can in fact be less than $1/2$! This occurs for example when $X \sim \phi_{0,1}$ and $Z \sim \phi_{0.1,0.5}$ (see Fig. 1); here the nearest neighbor rule when Y is drawn from $\phi_{0,1}$ is correct only around 44.5% of the time! i.e., if you draw once from $\phi_{0,1}$ and once from $\phi_{0.1,0.5}$, a second draw from $\phi_{0,1}$ will—more than half the time—be closer to the point from the other distribution. However, if $\epsilon \neq 0$, then Eq. 4 turns out to still be true in the Gaussian setting

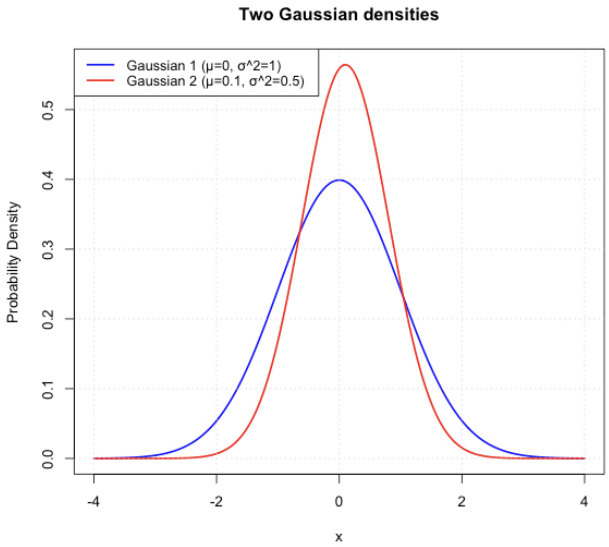

Figure 1: *Two Gaussian densities which do not satisfy a condition used in the proof with equal variance .*

with (possibly) different variances (if one of the two probabilities is less than $1/2$, the other

almost magically compensates so that their sum is greater than 1), giving us a more general result.

Theorem IV.1 *Suppose that $X \sim \phi_{\mu_X,\sigma_X^2}$, $Z \sim \phi_{\mu_X+\epsilon,\sigma_Z^2}$, and that Y is a 50–50 mixture of the two distributions, where $\mu_X \in \mathbb{R}$, $\epsilon \neq 0$, and $\sigma_X^2 \neq \sigma_Z^2$ are all unknown. Suppose we have x, z, and y generated respectively from these three distributions. Then the decision rule,*

$$C^{dist}(y) := \begin{cases} \phi_{\mu_X,\sigma_X^2} & \text{if } |x-y| < |z-y| \\ \phi_{\mu_X+\epsilon,\sigma_Z^2} & \text{if } |x-y| > |z-y|, \end{cases} \tag{9}$$

has a probability of being correct greater than $1/2$.

Outline of the proof. This proof involves two steps near the end which lack rigor; these are highlighted in the text. Without loss of generality, rewrite σ_Z^2 as $\beta\sigma_X^2$ where β is some unknown positive number (since σ_X^2 and σ_Z^2 are unknown). As before, we first calculate

$$P^* = \mathbb{P}(|X-Y| < |Z-Y| \mid Y \sim \phi_{\mu_X,\sigma_X^2})$$

by separating the calculation into two terms P_1^* and P_2^* as in Theorem III.1. P_1^* is now:

$$P_1^* = 2\int_{r=-\infty}^{\infty}\int_{\alpha=0}^{\infty} \Phi_{0,1}\left(\frac{r}{\sigma_X}\right) \frac{1}{\sqrt{2\pi}\sqrt{\beta}\,\sigma_X} e^{-\frac{1}{2\beta\sigma_X^2}(r+\alpha-\epsilon)^2} \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2\sigma_X^2}(r-\alpha)^2} d\alpha dr. \tag{10}$$

We then merge the two Gaussian pdfs into a constant (w.r.t. α) and another Gaussian pdf (a function of α), and integrate out the latter as before, which gives—after copious algebra, that:

$$P_1^* := 2\int_{r=-\infty}^{\infty} \Phi_{0,1}\left(\frac{r}{\sigma_X}\right) \frac{1}{\sqrt{2\pi}\sqrt{1+\beta}\,\sigma_X} \exp\left\{-\frac{1}{2(1+\beta)\sigma_X^2}(2r-\epsilon)^2\right\} \times \\ \times \Phi_{0,1}\left(\frac{\beta-1}{\sqrt{1+\beta}\sqrt{\beta}\,\sigma_X} r + \frac{1}{\sqrt{1+\beta}\sqrt{\beta}\,\sigma_X}\epsilon\right) dr.$$

Due to a result we will shortly use, it will be useful to get the exponential term—which is more or less a Gaussian pdf—in terms of r and not $2r$. After some algebra, we find that:

$$\frac{1}{\sqrt{2\pi}\sqrt{1+\beta}\,\sigma_X} \exp\left\{-\frac{1}{2(1+\beta)\sigma_X^2}(2r-\epsilon)^2\right\} = \frac{1}{2\sigma*}\phi_{0,1}\left(\frac{r-\frac{\epsilon}{2}}{\sigma*}\right),$$

where

$$\sigma* = \frac{1+\beta}{2}\sigma_X^2\,.$$

To continue, we recall a result from Owen ([3], pg. 407):

$$\int_{m=-\infty}^{\infty} \Phi_{0,1}(a+bm)\Phi_{0,1}(c+dm)\phi_{0,1}(m)dm = \tag{11}$$
$$\frac{1}{2}\Phi_{0,1}\left(\frac{a}{\sqrt{1+b^2}}\right) + \frac{1}{2}\Phi_{0,1}\left(\frac{c}{\sqrt{1+d^2}}\right) - T\left(\frac{a}{\sqrt{1+b^2}}, \frac{c+cb^2-abd}{a\sqrt{1+b^2+d^2}}\right) \\ - T\left(\frac{c}{\sqrt{1+d^2}}, \frac{a+ad^2-bcd}{c\sqrt{1+b^2+d^2}}\right),$$

where T is Owen's T function (see [2]):

$$T(u,v) = \frac{1}{2\pi}\int_{t=0}^{v} \frac{e^{-\frac{1}{2}u^2}(1+t^2)}{1+t^2}dt. \qquad (-\infty < u,\, v < +\infty)$$

In order to invoke this result, we make the change of variable $m = (r - \epsilon/2)/\sigma*$ and then calculate a, b, c, and d in our setting, obtaining $a = \epsilon/(2\sigma_X)$, $b = \sqrt{1+\beta}/2$, $c = (\sqrt{1+\beta}\,\epsilon)/(2\sqrt{\beta}\,\sigma_X)$, and $d = (\beta-1)/(2\sqrt{\beta})$. Note that if ac were not positive (as it clearly is in our case), there is an extra term in Eq. 11 which we have not shown for clarity. Plugging these into Eq. 11, we get that:

$$P_1^* = \frac{1}{2}\Phi_{0,1}\left(\frac{\epsilon}{\sqrt{5+\beta}\,\sigma_X}\right) + \frac{1}{2}\Phi_{0,1}\left(\frac{\epsilon}{\sqrt{1+\beta}\,\sigma_X}\right) - T\left(\frac{\epsilon}{\sqrt{5+\beta}\,\sigma_X}, \frac{3}{\sqrt{1+2\beta}}\right) \qquad (12)$$

$$- T\left(\frac{\epsilon}{\sqrt{1+\beta}\,\sigma_X}, \frac{1}{\sqrt{1+2\beta}}\right). \qquad (13)$$

One must then run through essentially the same calculations for P_2^*, and the result is that P_2^* gives the same result as P_1^* except that ϵ is replaced by $-\epsilon$. To finally calculate P^* itself, we first note two things: (i) $\Phi_{0,1}(\delta) = 1 - \Phi_{0,1}(-\delta)$ for any $\delta \in \mathbb{R}$; (ii) Owen's T function satisfies $T(-u,v) = T(u,v)$. Using these facts, we obtain:

$$P^* = P_1^* + P_2^* = 1 - 2T\left(\frac{\epsilon}{\sqrt{5+\beta}\,\sigma_X}, \frac{3}{\sqrt{1+2\beta}}\right) - 2T\left(\frac{\epsilon}{\sqrt{1+\beta}\,\sigma_X}, \frac{1}{\sqrt{1+2\beta}}\right).$$

Since we cannot count on symmetry here, we now have to perform this whole process again to calculate the other term:

$$P^{**} = \mathbb{P}(|X-Y| > |Z-Y| \mid Y \sim \phi_{\mu_{X+\epsilon},\sigma_Z^2}).$$

Recall that σ_Z^2 is still equal to $\beta\sigma_X^2$ and that this is the *same* fixed unknown β as we have just worked with. Thus it is also true that $\sigma_X^2 = (1/\beta)\cdot\sigma_Z^2$. It turns out that the solution for P^{**} takes exactly the same form as that of P^* except that β is replaced—wherever it is found—by $1/\beta$, and σ_X by σ_Z. We then revert the σ_Z in P^{**} back to σ_X by multiplying by $1/\sqrt{\beta}$, and thus obtain:

$$P^{**} = 1 - 2T\left(\frac{\epsilon}{\sqrt{1+5\beta}\,\sigma_X}, \frac{3}{\sqrt{1+2/\beta}}\right) - 2T\left(\frac{\epsilon}{\sqrt{1+\beta}\,\sigma_X}, \frac{1}{\sqrt{1+2/\beta}}\right).$$

The final probability we are looking for, which is a function of ϵ and β, is given by $P(\epsilon,\beta) = (1/2)\cdot P^*(\epsilon,\beta) + (1/2)\cdot P^{**}(\epsilon,\beta)$ and we must prove that this is greater than $1/2$ if $\epsilon \neq 0$ and $\sigma_X^2 \neq \sigma_Z^2$. Let us denote by $\mathcal{T}(\beta,\epsilon)$ the sum of the four Owen T function integrals (ignoring the negative sign and putting aside the factor of $1/2\pi$ in front of each of them). The result will then be true if $\mathcal{T} < \pi$. Notice that each of the four Owen T function integrals is of the form:

$$\int_0^{f(\beta)} \frac{e^{-g(\beta)\epsilon^2(1+t^2)}}{1+t^2}dt,$$

where f and g output only positive numbers. For any $\beta > 0$, the largest this integral can get is when $\epsilon = 0$, but since $\epsilon \neq 0$, it is more precise to say that as $\epsilon \to 0$, this integral is monotically increasing for fixed $\beta > 0$.

If we want to bound $\mathcal{T}$ from above, we can simply bound each of the four integrals from above by setting $\epsilon = 0$ (lack of rigor #1). Each of the four integrals then takes the more simpler form

$$\int_0^{f(\beta)} \frac{1}{1+t^2} dt,$$

which is in fact exactly $\operatorname{atan}(f(\beta))$. Thus:

$$\mathcal{T}(\beta, 0) = \operatorname{atan}\left(\frac{3}{\sqrt{1+2\beta}}\right) + \operatorname{atan}\left(\frac{1}{\sqrt{1+2\beta}}\right) + \operatorname{atan}\left(\frac{3}{\sqrt{1+2/\beta}}\right) + \operatorname{atan}\left(\frac{1}{\sqrt{1+2/\beta}}\right).$$

This is a fairly nasty function to maximize analytically (lack of rigor #2). Symbolic differentiation and root finding using Wolfram Alpha shows (or if you like, suggests) that this function $\mathcal{T}(\beta, 0)$ has a unique maximum at $\beta = 1$, i.e., when $\sigma_X^2 = \sigma_Z^2$, which is the excluded case in the theorem's statement. Consequently, for any $\sigma_X^2 \neq \sigma_Z^2$, $\mathcal{T}(\beta, 0) < \mathcal{T}(1, 0) = 2\operatorname{atan}(\sqrt{3}) + 2\operatorname{atan}(1/\sqrt{3}) = \pi$ by elementary properties of the atan function. Thus $P > (1/2\pi) \cdot \pi = 1/2$ and the result is proved. □

Remark IV.1 *The question of the optimality of this nearest-neighbor rule under these conditions remains entirely open.*

V. More general cases

(Below are some notes on more general cases, without proofs and potentially with errors.)

We can now ask whether this kind of result can be extend to other densities or distributions than Gaussian ones. For example, is this result true in general if f_X and f_Z have probability density functions? The following page of calculations leads to Conjecture V.1 below. To try and take a step in the direction of an answer, we first note that the equation to be proved (Eq. 4) can be rewritten:

$$\mathbb{P}(|X - X'| < |Z - X'|) + \mathbb{P}(|X^* - Z'| > |Z^* - Z'|) > 1, \tag{14}$$

where X, X', and X^* are independent variables each with density f_X, and Z, Z', and Z^* independent variables each with density f_Z. We then remark that since

$$\mathbb{P}(|X - X'| < |Z - X'|) + \mathbb{P}(|X - X'| > |Z - X'|) + \mathbb{P}(|X^* - Z'| > |Z^* - Z'|) + \mathbb{P}(|X^* - Z'| < |Z^* - Z'|) = 2,$$

Eq. 14 will be true if and only if

$$\mathbb{P}(|X - X'| > |Z - X'|) + \mathbb{P}(|X^* - Z'| < |Z^* - Z'|) \leq 1. \tag{15}$$

Note that the left-hand side of Eq. 14 can be rewritten as $\mathbb{E}[W_1]$, where

$$W_1 = \mathbb{1}_{|X-X'|<|Z-X'|} + \mathbb{1}_{|X^*-Z'|>|Z^*-Z'|}$$

is a random variable that can take the values 0, 1, or 2. Similarly, the left-hand side of Eq. 15 can be rewritten as $\mathbb{E}[W_2]$, where

$$W_2 = \mathbb{1}_{|X-X'|>|Z-X'|} + \mathbb{1}_{|X^*-Z'|<|Z^*-Z'|}$$

is also a random variable that can take the values 0, 1, or 2. Thus the statement we wish to prove will be true if and only if $\mathbb{E}[W_2] \leq \mathbb{E}[W_1]$. By writing out these two expectations in the form:

$$\mathbb{E}[W] = 0 \cdot \mathbb{P}[W = 0] + 1 \cdot \mathbb{P}[W = 1] + 2 \cdot \mathbb{P}[W = 2]$$

and using independence, we quickly see that $\mathbb{E}[W_2] \leq \mathbb{E}[W_1]$ is equivalent to it being more likely that points X' and Z' are both closer to the other generated point from their *own* distribution than both being closer to the generated point from the *other* distribution. Or, to put it mathematically, $\mathbb{E}[W_2] \leq \mathbb{E}[W_1]$ if and only if:

$$\mathbb{P}(|X - X'| > |Z - X'|) \cdot \mathbb{P}(|X^* - Z'| < |Z^* - Z'|) \leq \mathbb{P}(|X - X'| < |Z - X'|) \cdot \mathbb{P}(|X^* - Z'| > |Z^* - Z'|). \tag{16}$$

Each of the four probabilities in this expression can be expanded as triple integrals and then simplified into double integrals using the same kind of steps as in the proofs of Theorems III.1 and IV.1. For instance,

$$\begin{aligned}\mathbb{P}(|X - X'| > |Z - X'|) &= \int_{x=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \int_{x'=-\infty}^{\infty} \mathbb{1}_{\{|x-x'|>|z-x'|\}} f_X(x) f_Z(z) f_X(x') dx dz dx' \\ &= \int_{x=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \left[\left(1 - F_X\left(\frac{x+z}{2}\right)\right) \mathbb{1}_{\{x<z\}} + F_X\left(\frac{x+z}{2}\right) \mathbb{1}_{\{x>z\}} \right] f_X(x) f_Z(z) dx dz,\end{aligned}$$

where F refers to the cdf of the referenced variable. If you perform this expansion for each of the four probabilities in Eq. 16 and then multiply out and do some fun algebra, many terms cancel, and it turns out that the result we wish to prove overall will be true if and only if the following double integral is non-negative:

$$\int_{r=-\infty}^{\infty} \int_{\alpha=0}^{\infty} \left(F_X(r) - F_Z(r)\right) \left(f_X(r-\alpha) f_Z(r+\alpha) - f_X(r+\alpha) f_Z(r-\alpha)\right) d\alpha dr \geq 0.$$

Let us therefore state this as a conjecture.

Conjecture V.1 *Let $X \sim f_X$ and $Z \sim f_Z$ where f_X and f_Z are densities with cdfs F_X and F_Z respectively, and X and Z are independent. Then:*

$$\int_{r=-\infty}^{\infty} \int_{\alpha=0}^{\infty} \left(F_X(r) - F_Z(r)\right) \left(f_X(r-\alpha) f_Z(r+\alpha) - f_X(r+\alpha) f_Z(r-\alpha)\right) d\alpha dr \geq 0.$$

Remark V.1 *Currently we have made no progress on proving or disproving this statement of the problem. One gets the feeling that if for a given r, $F_X(r) > F_Z(r)$ (i.e., $F_X(r) - F_Z(r)$ is a positive number), then f_X has more density "to the left" of r than f_Z and consequently we could expect that—more often than not—the integral over positive α of $f_X(r-\alpha) f_Z(r+\alpha) - f_X(r+\alpha) f_Z(r-\alpha)$ would also be positive, and vice versa if $F_X(r) < F_Z(r)$, leading to on average a positive double integral, but this is not much of an argument.*

Remark V.2 *Some possible routes to investigate this double integral:*

- *The Wasserstein-1 distance?*
- *Convolutions?*
- *Rewriting the cdfs as integrals of pdfs?*

Remark V.3 *If this conjecture is true, the question is then whether the result of interest is also true for general probability distributions and not simply densities.*

References

[1] Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162), 2019.

[2] Donald B Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090, 1956.

[3] Donald Bruce Owen. A table of normal integrals: A table. *Communications in Statistics-Simulation and Computation*, 9(4):389–419, 1980.

[4] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.